The experience of analyzing the digital track of fans to assess the holding of mass events: the example of the 2018 FIFA World Cup

Lyubov Dmitrievna Zabokritskaya
Department of sociology and technology of public administration
Ural Federation University
Yekaterinburg, Russia
zabokritskaya@urfu.ru

Alina Vladimirovna Kulminskaia
Department of sociology and technology of public administration
Ural Federation University
Yekaterinburg, Russia
a.v.kulminskaia@urfu.ru

Evgeniy Igorevich Komotsky
Department of System Analysis and Decision Making
Ural Federation University
Yekaterinburg, Russia
evgeny.komotsky@urfu.ru

Abstract **– The authors explore the Russian and international experience of studying the digital footprint in this article. In particular, it analyzed the digital trace of the fans of the World Cup in 2018. The basis of the research was selected messages on the social network Twitter, marked the official hash tags of the event. As a result of the study, it turned out to measure the psycho-emotional state of the fans, to assess the general emotional background, which accompanied the championship, to single out the periods of the most intensive communication. Together with other search dedicated to the 2018 FIFA World Cup, this research will assess the atmosphere in which the FIFA World Cup was held. In the course of the work, the author's methodology for studying the digital trace was created and tested. The method proposed by the authors can be used to assess the social mood of citizens, and it is also applicable to Big data. It was found that the advantage of the presented technique is high speed and large volumes of analyzed data.**

Keywords – digital trace, Twitter, Big data, interdisciplinary research, the 2018 FIFA World Cup.

I. Introduction

Holding a major international event is a serious incentive for the development of all subsystems of the state. Understanding the threats and opportunities that accompany the process of the organizing the World Cup began long before it was held and will continue for some time. The quality assessment of the championship includes such indicators as the degree of readiness of the transport system to transport a record-high number of passengers, the state of the hotel fund and the staff's ability to meet foreign guests, the effectiveness of the system of medical care in crowded places, the consistency of public utilities, support for communication quality, organization of policing and so on. Specialists in each field are now engaged in analyzing experience and making recommendations for improving work in their professional field. In turn, we set ourselves the task of assessing the psycho-emotional state of the fans by analyzing their digital footprint on social networks. In conjunction with other search on the 2018 FIFA World Cup, our research will assess the atmosphere in which the FIFA World Cup was held.

The research of the psycho-emotional state of the fans during the days of the 2018 FIFA World Cup was carried out by fixing a digital track on the social network Twitter. The fixing of the digital trace was carried out using the technique developed by the authors. The results of the research showed: firstly, the high efficiency of the proposed method of analyzing the digital trace of the fans, and secondly, the high information content of the data obtained. The results will be useful for a comprehensive assessment of the socio-economic impact of the World Cup in the Russian Federation.

II. Literature Review

The key concept of our research is “Digital Footprint”. In the literature this term is adjacent to the concept of “digital shadow”, which is much broader and includes more ways of collecting information about the population [1]. The concept of digital is tied to direct use in the Internet space. The term “digital footprint” is a word-coinage inspired by the model of the English (“digital footprint”). Such a narrow definition is also important for us because the footprint always has a clearer reflection of the essence than the shadow, which by its nature is changeable, has fuzzy outlines and is a symbol of uncertainty. Digital footprint provides an opportunity to create a fairly clear picture of each individual Internet user, and when using a combination of data, a group of users, united by some parameters, allows you to create a comprehensive picture of the problem under study.

In our research under the digital footprint we understand the result of the demonstrative behavior of Internet users, which is fixed in digital space in text or visual form, the purpose of which is to identify the user. Thus, the presentation function of social networks is realized [2].

If we look at the term's history then the first publications about the digital trace were made by Western IT specialists, but they primarily concern the mechanisms of the formation of the trace and data collection technologies. In Russia due to the specific trends of changes in the modern legal field, the analysis of the digital trace in forensic science is developing most actively. This issue is dealt with Meshcheryakov V.A., Smushkin V.A., Lushin E.A. and etc. These authors first of all try to concretize the content of the term in order to increase its law enforcement. In this cluster of sources, the materiality of the digital trace (materiality, the ability to equate the trace to the consequences of the actual behavioral practice) and its forensic (evidentiary) significance are emphasized [3]. This group of research is interesting because they reinforce the materiality of the digital footprint: virtual practices go off-line and influence non-virtual practices. Regarding the disclosure of the mechanisms for the formation of the digital footprint, psychologists and sociologists are actively working: K. Feher, S. Garfinkel, D. Cox, Tulupyeva T.V, Tafintseva A.S, Andreev I.L, Nazarova L.N. This group of scientists consider the digital footprint as part of public (often - demonstrative) behavior, a way to form and maintain self-identification. Thus, we can see that the digital footprint of the individual user can be multidimensional: each social network has its own specific content and usage. For example, VKontakte is more entertaining than Facebook, which in recent years has become a platform for professional communication and socially significant discussions. Taken by us to study the social network Twitter is a convenient platform for the expression of a momentary psycho-emotional state. In particular, Chizhikov A.V. notes the possibility of searching social attitudes by analyzing the emotional coloring of posts on Twitter [4]. McCormick and others noted that Twitter is the largest observation's source of the human behavior to date [5].

Domestic experts in the field of management are still quite inattentive to the possibilities of analyzing the digital trace. There are separate articles devoted, for example, to the analysis of a student's digital footprint (Stepanenko A.V., Feschenko A.V.) or the search for a mass customization's method of clothing is based on the analysis of the digital footprint (N. Sakharova). This kind of search is aimed at finding instruments of influence on a specific socio-demographic group through the study of demonstrative behavior in social networks. But, again, such research is single, which indicates a weak development of the technology of the digital trace analysis in management in the Russian Federation.

If we talk about the international scientific community, then the situation is radically different. The relevance of studying the digital footprint, in the first place, was realized by marketers to search for the target audience, as well as economists [6]. So the banking sector employees use the digital footprint as one of the sources of information about their customers. On the basis of such data, in particular, a decision is made about the possibility of issuing a loan to an individual.

According to many researchers the ability to aggregate a huge number of digital traces of human behavior through media platforms is a new paradigm of data collection [7]. Information about the digital wake is actively used in all socio-economic spheres of human life from analyzing political processes [8] to analyzing information in the field of medicine [9]. We believe that Russia has sufficient research and development potential to conduct a qualitative analysis of the digital trace of users of social networks.

III. RESEARCH

Twitter was chosen as a social network to collect data on the psycho-emotional state of the fans [10]. It is Twitter that is the main online space for citizens, where you can publicly express your reaction to events and, therefore, the source of data for the social sciences. In addition, Twitter is an international social network in which you can explore fans from different countries. The practical purpose of the search was to test the automated method of analyzing messages for their emotional coloring (for example, the reactions of fans in the days of FIFA 2018). It was important for us to make sure that the method proposed by the authors can be used to assess the social mood of citizens, and also applicable to Big data. The presented research is a pilot project. In total, over the period from June 19 to July 10, 2018, more than 1 million messages were unloaded containing official hashtags of the championship. Messages were analyzed in the five most common languages: Russian, English, French, Spanish and Portuguese.

It should be noted that several years ago the task of analyzing the emotional background was a very complicated and time-consuming process, since that it has been about processing large volumes of information. Due the development of machine learning technologies and, in particular, the GloVe and word2vec models, it has become possible to accurately determine the emotional background in different languages.

Since, in essence, the proposed method has been an automated version of content analysis, then at the first stage of the research it was necessary to determine a list of keywords expressing the essence of "FIFA 2018", for example: "Russia2018", "FIFA", "Mondial", etc. To form a set of key words, we iteratively unloaded tweets, and then for each language, we extracted those keywords that were typical for the messages of fans using the TF-IDF algorithm. The selected key words were loaded into the TextBlob library.

TextBlob is a library for processing text data. TextBlob is a simple API for immersion into common natural language processing (NLP) tasks, and performs such tasks as: marking up parts of speech, extracting keywords from text, analyzing key, classifying, translating, and etc. This library is intended for the Python programming language. Initially, the use of the TextBlob library allowed us to carry out the selection of synonyms, the translation of key words into all the languages of the library, and the correction of the spelling.

Then, using the Python code, every day from June 19 to July 10, a collection of messages from Twitter was collected. The authors note also that Python can be used to get additional information about the post [“Meet Python, machine learning and the NLTK library” https://www.ibm.com/developerworks/ru/library/os-pythonnltk/]. For the theoretical task of our research (the assessment of the psycho-emotional state of the fans) it was enough to receive only the messages themselves (messages containing key words), as well as information about the date of their placement and language.

Our hypothesis was that the emotionality of the fans will gradually increase to the finals of the 2018 World Cup, and the emotionality of the statements will depend directly on the results of the match. Thus, the tone of the statements should depend on the language of the fans and be relevant to the outcome of the matches.

After July 10th, the entire collection of messages collected during the championship days was subjected to engine analysis. Through the TextBlob library, an analysis was made of the tonality of each of the messages as a whole. The tone of the message was estimated from 1 to -1. The most positive messages = 1, and the most negative messages were assigned the value -1. Messages that are neutral are = 0.

In addition to solving the problem of the determining the polarity of the message as a whole, such tasks as:

-selection of key information objects (people, places, the main subject of the message, etc.) - for this, the Named Entity Recognition functionality from the NLTK and TextBlob libraries for Python was used;

- recognition of emotions (joy, anger, sadness, fear, disgust) for different languages - for this, a classifier trained by Naïve Bayes was used based on the following data sets (EmoBank, ISEAR, SemEval) from the library scikit-learn for Python.

IV. RESULTS

Based on the results of uploading more than 1 million messages, which contained the official hashtags of the championship and determining the tone of these messages, the authors studied the psycho-emotional state of the fans during the world Cup. Positive messages were recognized messages with a tone level from 0.5 to 1, negative from -0.5 to -1. Neutral messages ranged from 0.49 to -0.49.

The data presented in Figure 1 shows that the psycho-emotional state of the fans during the days of the World Cup tended to increase from the beginning of the Championship to the final. In general, the positive mood of the fans, significantly prevailed over negative emotions The number of positive or negative messages prevails over neutral ones, which indicates a high level of emotionality of these messages as a whole.

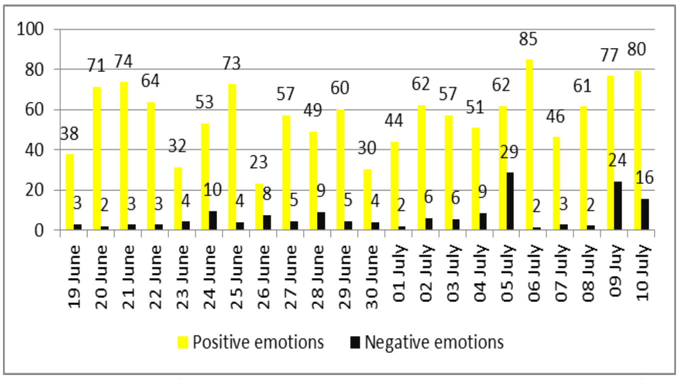

Fig. 1. The share of positive and negative messages with the hashtag of the championship in Twitter

In fig. 1% of the total amount of messages does not give 100%, since the rest are made up of messages that have a neutral emotional tone.

Bursts of emotional stress are directly dependent on the matches, which took place on the Championship days and their results. Then we measured the integral representation of the general emotional background for each language during the 2018 World Cup, namely Russian, English, Portuguese, French and Spanish.

In Figure 2 you can see that the most positive users are Russian-speaking, and the most negative ones are French-speaking. At the same time, it is frankly speaking fans, who are generally more emotional than fans speaking other languages. Perhaps it was such a high emotional support of the fans, which helped the French team to win the 2018 World Cup.

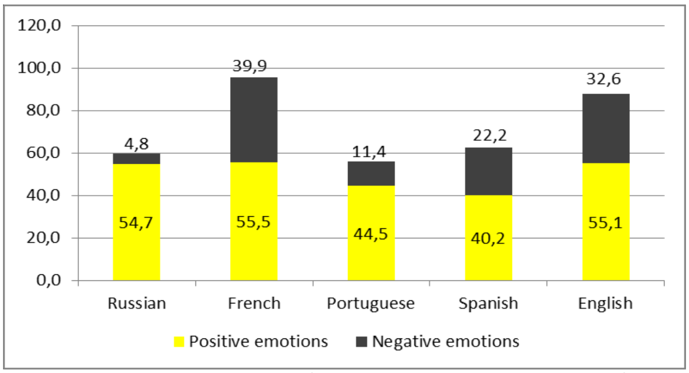

Fig. 2. An integral representation of the general emotional background for the main languages of the World Cup

The data presented shows the dynamics of the psycho-emotional state of the fans by the days of the World Cup, as well as the general emotionality of the fans speaking the main FIFA-2018 languages. It was also important for the authors to determine the specific emotional background emanating from the event. For this purpose, through the classifier trained by Naïve Bayes, from the library scikit-learn for the Python language, messages from Twitter were classified into five groups: anger, fear, gladness, sadness, disgust (Table 1).

TABLE I. DOMINATING EMOTIONS IN EACH LANGUAGE GROUP DURING THE 2018 WORLD CUP IN RUSSIA

Type of emotion	***Russian***	***English***	***Spanish***	***Portuguese***	***French***
sadness	18,2	18,3	*19,3*	18,1	14,2
gladness	18,2	*34,2*	26,8	26,0	*29,0*
fear	10,0	8,7	*11,5*	10,4	7,3
disgust	10,6	10,1	*12,6*	10,5	6,7
anger	*28,8*	10,0	9,4	9,8	7,1
Total emotionality	85,8	81,2	79,6	74,8	64,2

a* The sum of per cent does not give 100%, the remaining messages were neutral

The data presented in Table 1 allows you to see the total distribution of emotions reflected in the messages on the social network Twitter for each language, as well as the dominant emotion in each language group during the 2018 World Cup in Russia. So, for example, for Russian-speaking users, anger mostly prevailed in messages (28.8%), for other users the distribution of emotions, English and French fans showed a greater degree of joy (34.2% and 29%, respectively). Spanish fans showed more often such emotions as sadness (19.3%), fear (11.5%) and disgust (12.6%). For example, the number of messages with such types of emotional coloring prevailed among Spanish-speaking fans immediately after the unexpected loss of the Spanish national team in the 1/8 of the Russian national team.

IV. DISCUSSION OF THE RESULTS

In the course of the search our hypothesis was fully confirmed. Indeed, the fans' emotionality grew towards the finals of the 2018 FIFA World Cup. The emotionality of the statements of messages emanating from the fans depended directly on the results of the Championship matches. In general, on the days of the 2018 World Cup, the positive psycho-emotional mood of the fans dominated, which also testifies to the quality of the organization and conduct of the entire 2018 World Cup.

Thus, the study of the psycho-emotional state of fans through the analysis of their digital footprint in social networks showed a high level of psycho-emotional stress, as well as a significant prevalence of positive messages over negative messages.

In addition, we tested an automated method for analyzing messages for their emotional coloring. The method proposed by the authors can be used to assess the social mood of citizens, and it is also applicable to Big data. It was found that the advantage of the presented technique is high speed and large volumes of analyzed data. Since the purpose of the study was not just to study the messages for their emotional coloration, but rather to test the automated Big data technique, the sample was representative. The use of an automated message analysis technique allows the sample to be brought close to 100% and also to reduce the human factor in the collection, processing and analysis of the results obtained. The disadvantages of the proposed method is the lack of transparency regarding the selection of tweets in the stream, with the exception of the language of social network users.

V. CONCLUSIONS

In general, the experience of studying digital trace can be considered successful. Our research opens up prospects for other work in the field of Big data and Human digital. At the same time, during the search, some shortcomings of the technique were revealed. In particular, in the tweets selected for the tags given by us, it was a question directly of the Championship, players, successes and failures of the teams. In order to assess the country's image or the effectiveness of the work of the individual services and subsystems of the state, this is not enough. This problem can be solved in two ways. The first way is to search for tags within a language group that display messages containing information of interest to us. This path is vulnerable, since that it has been possible that the necessary tweets are not tagged at all. Therefore, the second way is preferable, when planning a large event, create tags of several different meanings. For example, it has not been just # fifa2018, but #russiancity or #russianfood, i.e. tags that can, on the one hand, help to identify a topic, and on the other, reveal more topics for guests to publish. This should be thought out at the planning stage of the event and include the promotion of tags in the promotion strategy.

The second point that requires attention is the lack of the method for determining the formation's causes of the dominant emotion. Here one can assume that the dominant emotion is a reflection of the national mentality, refracted through the prism of success or failure of the own football team. At the same time, the absence of an obvious negative emotional background allows us to speak about a favorable emotional atmosphere during the matches. Despite the identified shortcomings, the use of Textblob seems promising to us in solving managerial problems, which include the assessment of the quality of major events. The proposed method is able to streamline the process of working with large volumes of information coming from social networks. This is especially important due to the fact that the disordered nature of working with information, the lack of necessary skills and tools, as well as cognitive distortions associated with the bias in assessing current events, which are called the key risk of the innovation in modern society.

In general, we see the scientific community's interest in this kind of the research. The concept of "digital footprint" is becoming widely known. Many people are moving towards the conscious formation of their digital footprint. In the future, this will fundamentally change the research approach: from the analysis of spontaneous behavior to the analysis of the behavior of the selective and it will raise the ethics' question of this kind of the research. All these changes open up a broad field of the interdisciplinary cooperation.

References

[1] L.A. Boyarkina, V.V. Boyarkin (2016) Tsifrovoy sled i tsifrovaya ten kak proizvodnyye personalnykh dannykh, *Sborniki konferentsiy NITs sotsiosfera*, vol. 62, pp. 78-81.

[2] E. Goroshko (2011) Chirikayushchiy zhanr 2.0 Tvitter. Ili chto novogo poyavilos v virtualnom zhanrovedeni, *Vestnik Tverskogo gosudarstvennogo Universiteta,* №3, pp. 11-16.

[3] A.L. Osipenko (2009) Problemy vovlecheniya elektronno-tsifrovykh sledov v ugolovnyy protsess, *Nauchnyy vestnik Omskoy akademii MVD Rossii*, № 4(35), pp.31-34.

[4] A.V. Chizhik (2016*)* Faktory formirovaniya sotsialnogo nastroyeniya na osnove analiza emotsionalnoy okraski postov v russkoyazychnom Twitter, *Novyye informatsionnyye tekhnologii v avtomatizirovannykh sistemakh*, №19, pp. 61-64.

[5] TH. McCormick, H. Lee, N. Cesare, A. Shojaie, ES. Spiro (2017) Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing, *Sociological methods & research*, vol. 46 (3), pp. 390-421, DOI: 10.1177/0049124115605339.

[6] E. Cambria, HX. Wang, B. White (2014) Guest Editorial: Big Social Data Analysis, *Knowledge-based systems*, vol. 69, pp. 1-2, DOI: 10.1016/j.knosys.2014.07.002.

[7] B. Goncalves, N. Perra, A. Vespignani (2011) Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number, *Plos one*, vol. 6 (8), DOI: 10.1371/journal.pone.0022656.

[8] A. Hermida, SC. Lewis, R. Zamith (2009) Sourcing the Arab Spring: A Case Study of Andy Carvin's Sources on Twitter During the Tunisian and Egyptian Revolutions, *Journal of computer-mediated communication*, vol. 19 (3), pp. 479-499, DOI: 10.1111/jcc4.12074.

[9] C. Chew, G Eysenbach (2010) Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1, *Outbreak PLOS ONE*, vol. 5 (11), DOI: 10.1371/journal.pone.0014118.

[10] ML. Williams, P. Burnap, L. Sloan (2017) Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation, *Sociology-the journal of the British sociological association*, vol. 51 (6), pp. 1149-1168 DOI: 10.1177/0038038517708140.